Towards Interpretable, Data-derived Distributional Semantic Representations for Reasoning: A Dataset of Properties and Concepts

Pia Sommerauer, Antske Fokkens and Piek Vossen
Computational Lexicology and Terminology Lab
Vrije Universiteit Amsterdam
De Boelelaan 1105 Amsterdam, The Netherlands
pia.sommerauer@vu.nl, antske.fokkens@vu.nl, piek.vossen@vu.nl

Abstract

This paper proposes a framework for investigating which types of semantic properties are represented by distributional data. The core of our framework consists of relations between concepts and properties. We provide hypotheses on which properties are reflected in distributional data or not based on the type of relation. We outline strategies for creating a dataset of positive and negative examples for various semantic properties, which cannot easily be separated on the basis of general similarity (e.g. **fly**: *seagull*, *penguin*). This way, a distributional model can only distinguish between positive and negative examples through evidence for a target property. Once completed, this dataset can be used to test our hypotheses and work towards data-derived interpretable representations.

1 Introduction

When it comes to representations of word meaning, we currently have to choose between relatively transparent, interpretable representations that are low in coverage and opaque embedding representations with high coverage. While the former lend themselves well to reasoning, the latter are hard to interpret and their reasoning potential remains limited. Ideally, we would have 'the best of both worlds': data-derived, high-coverage transparent representations we can reason over. Reasoning over such vectors would open new opportunities for the study of phenomena at the core of lexical semantics, such as similarity and ambiguity (one form - multiple meanings) and variation (one meaning - multiple forms).

In this paper, we present a framework for analyzing what type of semantic information is present in distributional data as a first step towards such semantic representations. We consider word meaning from the perspective of semantic properties, which enables us to explain semantic similarity and dissimilarity and reason over word meanings. We propose a methodology that can be used to create datasets representing concepts and their semantic properties, which can be used to test hypotheses about what type of information is present in distributional models.

When trying to model the type of semantic information represented by linguistic context, the following questions arise: Which aspects about the meaning of a word can be expected to be mentioned in (written) utterances? Do people talk about the yellowness of lemons? Or would they rather give accounts of what lemons are used for? We propose a number of hypotheses about which type of semantic knowledge is encoded in the linguistic context based on the semantic relation between a particular concept and property.

If distributional vectors contain information about a semantic property, it should be possible to distinguish positive examples of the property from negative examples purely on the basis of the distributional vector. As distributional semantic representations usually provide good indications for general relatedness or similarity, one major pitfall of our approach is that words can easily be separated into positive and negative examples because they happen to fall into rather distinct categories. Therefore, we specifically aim to collect challenging examples (e.g. **fly**: *seagull*, *penguin* rather than **fly**: *seagull*, *table*). We propose a framework for sampling and defining concept-property pairs that, in future work, will be annotated and used to test our hypothesis. To the best of our knowledge, this will be the first dataset specifically designed to analyze the ability of embeddings to encode property information.

Besides being a diagnostic tool, we hope that

the resulting resource will provide complementary information to traditional lexical semantic representations. A core notion in lexical semantics is semantic similarity. Different lexical resources reflect this notion in different ways. Whereas Princeton Wordnet (Fellbaum, 2010; Miller, 1995) structures semantic knowledge in terms of hierarchical categories, we approach similarity from the perspective of property overlap. Implicitly, knowledge about property overlap is also represented in hierarchically structured categories, as they capture information about shared and distinguishing properties. We expect that the final dataset will be a complementary resource to WordNet as it could yield insights into semantic categorization in terms of semantic properties. Currently, our setup only takes English data in consideration, but we think that valuable insights could be gained from extending it to more languages thus enabling cross-linguistic comparisons.

The remainder of this paper is structured as follows: Section 2 outlines insights on semantic properties from various research domains. Based on this, we present a framework of properties and concepts in Section 3, followed by our method for creating our dataset suitable for testing our hypotheses in Section 4. We conclude and discuss the implications of our framework in Section 5.

2 Theoretical Background

This section provides an overview of theories and observations about the type of knowledge encoded in linguistic contexts. In general, we assume that semantic information can either be encoded explicitly (e.g. by expressions such as *lemons are yellow*) or implicitly (e.g. *the lemon rolled off the table*, which indirectly indicates that lemons have a round shape). Both sources of evidence provide sufficient information for humans to infer these properties. It is an open question to what extent this is represented by embedding models. We start this investigation by raising the question of what type of information is likely to be mentioned (either implicitly or explicitly) in natural language.

Different theoretical and applied fields have addressed this question, namely, language generation, corpus linguistics and cognitive theories of word meaning. We draw from approaches about referential expressions (Section 2.1), typical properties and concepts revealed in similes (Section 2.2) and afforded actions and processes (Section 2.3). The remainder of this section provides an outline of these factors which form the basis of our proposed framework, introduced in Section 3.

2.1 Gricean Maxims

One major function of language is to 'point' towards things in the world. This is explicitly modeled in approaches to natural language generation, which include referring expression generation (REG) as a subtask (Gatt and Krahmer, 2018). Dale and Reiter (1995)'s seminal work proposes to model REG in terms of Gricean maxims (Grice, 1975). In essence, humans are expected to refer to objects by being maximally informative while not providing more information than necessary, resulting in the use of maximally discriminative attributes. When given a choice of objects with a range of different, but partly overlapping attributes and the task of singling out a particular one, humans are expected to use only the attribute(s) which is (are) most informative.

Experimental data show that people do tend to overspecify (in as much as 50% of cases (Koolen et al., 2011)) for several reasons: Arts et al. (2011) argue that overspecification in terms of highly salient attributes may facilitate identification of the referent. Rubio-Fernández (2016) claim that the overspecification of color attributes can facilitate object search as it is easier to find something based on multiple pieces of information. For instance, finding a blue cup is easier if you can look for something blue and for a cup, in particular when the target object is the only cup *and* the only blue object. A complementary observation was made by Koolen et al. (2011), who show that overspecification increases with the difficulty of the reference task. However, color attributes also tend to be overspecified for objects which are typically described in terms of color, such as clothes. This later phenomenon possibly is language-dependent, as it was observed for English speakers but not Spanish speakers. More generally, Sedivy (2003) found that color attributes tend to be used redundantly for objects that have a high color-variability (i.e. things that naturally come in several colors, such as t-shirts). Complementary, Koolen et al. (2011) observe that overspecification occurs for concepts whose instances can be described in terms of many different attributes.

These insights have been obtained from highly controlled lab settings with limited situational

context. An attempt to generalize to the information included in utterances 'in the wild' can be seen as somewhat of a leap. Nevertheless, we expect that in general, people tend to avoid mentioning information which is already available to their interlocutors through their (physical) experience of the world. For instance, we expect that people would hardly ever specify the color or taste of a lemon (unless it is a highly unusual one), since this information is already available to people who have had some sort of experience with lemons. In contrast, we expect that people are more likely to specify target objects in terms of attributes (e.g. color) in case of high variability of attributes or in case strong association between concepts and attributes (typicality). The former could either be due to (1) the reference task being actually harder because of the high variety of attributes or (2) the observed tendency to overspecify in cases of high attribute-variability. The next section discusses how typicality can result in contexts that explicitly reflect shared knowledge.

2.2 Stereotypicality

Veale (2013) explores the way different semantic properties of concepts (most of which can be seen as having 'multifaceted' meanings) can be extracted from text corpora. He proposes that " [...] words[1] are represented as bundles of the typical properties and behaviors they are commonly shown to exhibit in everyday language" (Veale, 2013, p.1) and presents an automatic system to extract and reason over the different affective contents associated with concepts via their most salient properties. For instance, the word *baby* can receive a positive interpretation when appearing in a context highlighting cuteness and peacefulness, but just as well be used in less flattering descriptions such as *cry like a baby*.

Veale's approach shows that information about stereotypical concepts of a property is mentioned in natural language, as it relies on pattern extraction from corpora. Specifically, stereotype information tends to be expressed in similes of forms like *as* ADJECTIVE *as a* NOUN (e.g. *as mindless as a zombie*) or in the case of activities VERBing *like a* NOUN (e.g. *drooling like a zombie*) (Veale and Hao, 2007) .

It seems that implied information about concepts tends to be mentioned explicitly if the concept can serve as a particularly good example to illustrate the (implied) property. While it is unlikely to find instances stating the obvious (e.g. *coal is black*), it is more likely to find utterances in which the stereotypical concept is used to illustrate a property of something else (e.g. *eyes as black as coal*).

2.3 Common Actions and Affordances

Based on accounts in cognitive psychology and cognitive linguistics, we expect (highly implied) knowledge relating to specific types of afforded actions (as introduced by Gibson (1954)) likely to be reflected by linguistic context. Glenberg (1997) argues that a central component of our memory is a set of actions that are available to an agent in a certain situation, which he calls 'mesh'.

Glenberg and Robertson (2000) explore this notion by comparing embodied to high-dimensional (i.e. distributional) theories of meaning. Their experimental results indicate that distributional models provide good indications about the kinds of actions and processes concepts are usually involved in. They are, however, unable to reflect possible (i.e. *afforded*) actions that are highly unusual.

We hypothesize that this is due to a tendency of people to describe and report on specific events in the world, which consist of combinations of actions and processes. Specific events, in contrast to general properties, are very unlikely to be implied knowledge and therefore have to be communicated (e.g. *dogs have four legs* versus *My dog ran towards the ball*). A large corpus is more likely to contain patterns that arise from specific activities and processes (e.g. dogs will often be involved in running events), while unusual activities will be too erratic to lead to meaningful regularities in the data that end up represented in distributional models.

2.4 Summary of Factors

When determining whether a specific semantic property is likely to be encoded by distributional information, we consider the following factors to be relevant:

> **Impliedness**: Which information is already known, Which information has to be made explicit?
> **Variability**: Do the instances of a concept vary with respect to the target property?

[1] This paper is on word meaning. The expression 'word' should be read as referring to word meaning.

Typicality: Is a concept likely to be used to illustrate a property?

Affordedness: Do certain properties afford activities that instances of a concept engage in? In other words: Are there properties of a concept which enable certain activities?

3 Contextually Encoded Properties

Based on the observation outlined in Section 2, we present predictions about whether a specific distributional vector representation of a concept is likely to encode information about a specific semantic property or not. To operationalize this, we translate the factors discussed in Section 2 to descriptions of relations between concepts and properties. We assume that knowledge about properties of concepts is generally implied and hence unlikely to be expressed explicitly (following the Gricean maxim of quantity). However, there are a number of factors which cause violations against this general tendency. We translate these competing forces to relations between properties and concepts. We outline them below and summarize them in Table 1, which also provides an overview of our hypotheses.

Typicality. Typical properties of instances of a concept are usually also highly implied (e.g. *rose* - **red**). While a high level of impliedness in combination with Gricean maxims would mean that the property is unlikely to be mentioned explicitly, typicality may have the opposite effect. Based on the observations by Veale (2013), we expect that typical examples of a property can often serve to illustrate the property in a another concept (e.g. *coal* serves to illustrate blackness in the phrase *eyes as black as coal*, *rose* may serve to illustrate redness, etc). In contrast, properties that immediately come to mind when thinking of a concept, but not vice-versa are unlikely to be represented, but can be seen as highly implied (e.g. **green** is a typical property of *broccoli*, but *broccoli* is usually not used to illustrate greenness).

Affordedness. In general, we propose that afforded and usually performed activities are represented, while afforded and not usually performed activities are not (e.g. bowling ball - roll v.s. candle - roll). Usually performed activities can be seen as highly implied knowledge about a concept. However, the fact that activities usually form part of specific events (which are not part of our implied knowledge) makes them much more likely to be mentioned in communication than other highly implied properties. In addition to being afforded properties themselves, activities can also provide indirect evidence for other properties. In particular, they provide indirect evidence for those properties which enable the activity. For instance, *bowling balls* are commonly involved in rolling-activities. The context is likely to provide direct evidence of the activity **rolling** (e.g. *The bowling ball rolled by 5-foot-10*).[2] The same evidence can also serve as an indirect indication for the property affording the rolling-activity, namely being **round**. Many properties of a concept are, however, not necessarily reflected in activities. Consider, for instance *candles*: even though they are often round (an affording property for the activity of rolling), rolling is not something they typically do. In the remainder of this paper, we use the following sub-types of properties: We distinguish activities from attributes. Activities can be afforded and usually performed or afforded and not usually performed (or not afforded at all). Attributes can fall under any of the relations outlined here. In addition, they can afford activities.

Variability. This factor refers to the degree of variation in instances of a concept. In general, we propose that variable properties are likely to be represented by linguistic contexts because they can be relevant for further distinctions and are not automatically implied. For instance, a color attribute can distinguish between different subcategories of bears or distinguish between peppers with different tastes, knives can be used for different cooking activities or processes, etc. These variable properties can have different degrees of discriminatory power. On one end of the spectrum, they distinguish between different conceptual categories (e.g. subcategories of bears). At the other end of the spectrum, they distinguish instances of the same category (e.g. t-shirts of different colors or dogs trained for different activities). While in this later case, there is a very high probability of properties to be mentioned explicitly, we do not expect the evidence to be enough to be captured by a distributional semantic model: due to the high degree of variance, individual properties will be mentioned sporadically at best. Properties that can only apply to instances of concepts in exceptional cases are not expected to be represented.

[2] https://www.latimes.com/archives/la-xpm-1991-05-30-sp-3586-story.html

factor	present	absent
typicality	concept is typical of the property	property is typical of the concept
afforded activities	usually performed	possible but not usually performed
affording attributes	affording usually performed activities	not relevant for usually performed activities
variability (options)	limited (also values on a scale or opposites)	wide selection
variability (categories)	subcategories	not relevant for subcategories

Table 1: Overview of relations between concepts and properties: **present** and **absent** indicate whether the concept-property relation is hypothesized to be apparent from distributional data.

Table 1 provide an overview of the relevant factors and related prediction. A single concept-property pair can be related to more than one factor. For instance, *sky* - *blue* can be described in the following terms:

Implied : **blue** is a highly implied property of *sky*

Typical (concept) : **blue** is a typical property of *sky*

Typical (property) : *sky* is a stereotypical example of something which is **blue**

Variable (limited) : *skies* can also be **grey** or **black**

If at least one description falls under *present* in Table 1, we expect the context to contain evidence for the property. Whether this evidence is sufficient for a distributional model to represent the property is an open question.

4 A Dataset of Concepts and Properties

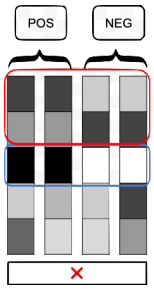
This section describes the design of our dataset. We first outline the experiments we envision, because they provide the motivation of some of the key properties of our dataset.

To conduct experiments on whether the predictions introduced in Section 3 hold, we plan to use approaches suggested in the field of investigating neural network representations, such as diagnostic classification (Belinkov et al., 2017; Hupkes et al., 2018; Derby et al., 2018). In particular, we plan to extend the experiments presented in Sommerauer and Fokkens (2018), which try to investigate whether dimensions of embedding representations can capture semantic properties. While this seems to be implied by the method of inferring the missing word in an analogy pair by means of vector subtraction and addition (Mikolov et al., 2013; Levy and Goldberg, 2014), analogy calculation methods have been heavily criticized, calling this notion into question (Linzen, 2016; Gladkova and Drozd, 2016; Gladkova et al., 2016). To shed light on this, we proposed an experimental set-up in which we tested whether a supervised machine learning system could successfully learn to distinguish vectors of words clearly associated with a property from vectors of words which are clearly not associated with the property.

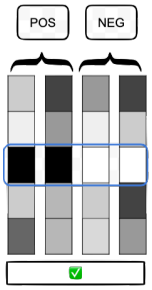
Any supervised classification approach relies on finding regularities which are shared among all or most examples of a particular class and distinguish them from other classes. Therefore, the distribution of positive and negative examples of properties is crucial to ensure that the vector dimensions discovered by the classifier actually correspond to the semantic property under investigation rather than some other information which happens to correlate with it. To illustrate the importance of the similarity distribution of positive and negative examples, consider the following: Suppose our dataset for the property **red** consists of names of red fruits (positive examples) and green garden plants (negative examples). If we train and test a classifier on such a dataset, it is very likely that it can reach relatively high performance. But did it learn to identify the semantic property **red** in a distribution? In such a case, it would be impossible to draw a clear conclusion for the following reasons: The names of the red fruits most likely share more properties than being red, such as having a sweet taste, being used for similar things, or largely falling into the category of berries. Consequently, more information connects these examples than the property **red**. The same holds for the negative examples: they belong to a relatively coherent category and probably share many properties. Many of these properties will not be shared with the positive examples. This means that a classifier can rely on a multitude of indications, none of which are necessarily evidence of the target property **red**. Figure 1 illustrates different scenarios of shared and distinguishing features.

To address this challenge, our dataset has to ad-

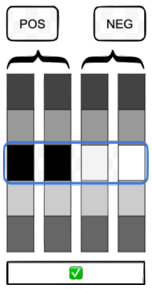
(accessed 2019/09/30)



(a) Target property (present in black, absent in white) and other dimensions correlates with positive and negative classes.



(b) The only dimension correlating with the positive and negative classes is the target dimension of the target property.



(c) Highly similar positive and negative examples that can only be distinguished by the dimension of the target property

Figure 1: A schematic representation of vectors of positive and negative examples of a property. To ensure that shared and distinguishing patterns identified by a classifier are representative of the target property, positive and negative examples should only be separable based on the target property.

here to the following requirements:

1. For each property, there is a sufficient number of positive and negative examples.
2. The distinction between positive and negative examples cannot be made on the basis of general similarity alone (see Figure 1). The candidates should include (1) positive examples that differ with respect to most properties except the target property, i.e. that have low overall similarity (e.g. **fly**: *seagull*, *airplane*) and (2) negative examples that share a number of properties with positive examples, creating high similarity between positive and negative examples (e.g. **fly**: *seagull*, *penguin*)

Most existing feature norm sets (McRae et al., 2005; Devereux et al., 2014) do not contain information about negative examples, as they only list (salient) properties of concepts. One might consider to derive negative examples by viewing all concepts *not* labeled with a certain feature as negative examples of the feature. This approach, however, results in a number of wrongly labeled instances, as positive cases are not always labeled as such (for instance, 18 out of 36 concepts labeled as **is_a_bird** are not labeled as **is_an_animal** in the CSLB feature norms (Devereux et al., 2014)).

Our main objective is to collect fine-grained information for property-concept pairs to fill this gap. Through crowd annotations, we aim to divide these property-concept pairs into three categories: Properties which apply to **all or most**, **some** or **hardly any or no** instances of a concept. We draw the line in the middle of the 'some' category, which encompasses different degrees of variability: while we expect attributes with little variance to have enough evidence for a model, attributes with a high degree of variability are most likely not encoded.

The second requirement can be fulfilled by controlling (a) the selection of target properties (see Section 4.1), (b) the selection of candidate concepts from resources (Section 4.2) and (c) the selection of particularly challenging examples in the distributional semantic space (Section 4.3). Sections 4.4 and 4.5 provide further details on the setup of our crowd sourcing task.

4.1 Selecting Challenging Properties

We select semantic properties which apply to concepts that are spread across traditional, taxonomic categories. We consider the following types of properties: perceptual attributes (e.g. colors, shapes, temperature), part attributes (e.g. *having wheels*), complex attributes (high level semantic categories such *being dangerous*) and activities (e.g. *swim*, *fly*). We hand-selected specific properties (listed in Section B of the Appendix) for each type based on the criteria of them cutting across taxonomic categories and applying to a large number of concepts.

4.2 Selecting Challenging Concepts

We collect candidate concepts from existing computational and psycholinguistic resources, listed in Table 2, and from a distributional model. By exploiting the feature norm sets and the stereotype data, we get a limited set of candidates 'for free' by searching for the selected properties directly.

By searching for target properties directly (e.g.

concepts associated with **round** in ConceptNet via the relations *HasProperty* or *NotHasProperty*), we only receive limited sets of examples, in particular with respect to negative candidates. Therefore, we extend the search by including concepts of particular traditional, taxonomic categories whose members we expect to have or not have the target property. We explain the idea through the activity **fly**.

Concepts that are similar and only differ with respect to **fly** or categories which contain positive and negative examples are particularly useful. We exploit this in our sampling strategy: we know that while most birds can fly, some cannot. The category of insects also contains both cases. In addition, we could add vehicles. While the first two categories contain similar concepts that share a large number of properties, the later category introduces words that share almost no properties with the first two except the target property.

For this type of search, we exploit the hyponymy relations of WordNet as well as properties from the feature norm sets and the corpus data. For WordNet, we manually select the synset representative of a category (based on synset members and definitions) and collect all lemmas of its hyponym-synsets. In the feature norm data, we simply search for the target property. In addition, we use the positive and negative examples derived from the CSLB norms and annotated by the crowd as described by Sommerauer and Fokkens (2018).

This strategy is successful for some properties (e.g. 105 probably positive and 256 probably negative candidates for the **black**) but less for others (e.g. 6 probably positive and 63 probably negative candidates for **round**). While we try to select negative examples that are difficult to distinguish from positive ones through other properties than the target property, it is not entirely clear whether this is the case. To extend our examples and at the same time target particularly challenging negative examples, we use an existing distributional model as a source of additional examples.

4.3 Challenging Examples using Embeddings

Distributional semantic models provide relatively good indications of word similarity, reflecting the assumption that words with similar meanings tend to appear in similar linguistic contexts. However, they cannot give us precise information about what makes words similar. The main challenge of our approach is to select examples that could not be distinguished purely on the basis of this distributional similarity. Therefore, we specifically select examples from a distributional model which have a very high chance of being classified wrongly based on their similarity (e.g. *penguin* for **fly**, or *heroine* for **dangerous** while other positive examples are weapons or animals). If it can be classified correctly, we can interpret this as good evidence for the property to be encoded in the distributional vector representation.

type	**resources**
feature norm sets	McRae et al. (2005), CSLB norms (Devereux et al., 2014)
lexicon	WordNet (Fellbaum, 2010; Miller, 1995) ConceptNet (Speer and Havasi, 2012)
stereotype data	concepts representing stereotypes of properties (Veale, 2013)
feature norms negative extension	subset annotated on top of the CSLB norms (Sommerauer and Fokkens, 2018), quantified McRae norms (Herbelot and Vecchi, 2015)

Table 2: Overview of resources.

To operationalize this, we select positive 'seed' words and calculate a vector representation for them by taking the average of the seeds. This allows us to specifically select candidates with embedding representations that are overall similar to positive examples of a property (by taking the n nearest neighbors of the averaged representation). We select these positive 'seeds' by using positive examples of a property we are confident about (i.e. we do not include concepts returned by a search for a category containing 'mixed' examples).

This results in a selection of candidate concepts which are very difficult to separate into positive and negative examples based on general similarity. We collect the 200 nearest neighbors of this approximate property representation. We exclude negative examples further away from the centroid than the furthest positive example by manual inspection. The embedding model used in this step is the skip-gram model with negative sampling (using recommended settings according to Levy et al. (2015)), trained on the full Wikipedia corpus (dump from August 2018).

4.4 Sampling for the Crowd

The strategies outlined above result in rather large numbers of candidates not all of which are useful

(e.g. the distributional model returns non-standard spelling variants and words other than nouns). We reduce and clean the resulting sets (1) by means of preprocessing and (2) sampling based on characteristics with potential impact on how well distributional data can represent information. The characteristics we consider are (1) different types of ambiguity, (2) psycholinguistic factors such as concreteness and familiarity represented in the MRC database (Coltheart, 1981), word frequency (3) the distance to the centroid vector calculated over all positive examples of a property.

type	**n syns**	**wup sim**	**min wup sim**	**cos syns**	**abs-conc**
homonyms	8.28 (6.97)	0.32 (0.16)	0.20 (0.18)	0.20 (0.20)	0.60
metaphors	8.32 (7.68)	0.35 (0.19)	0.24 (0.21)	0.24 (0.23)	0.45
metonymy (approx.)	3.01 (2.72)	0.53 (0.32)	0.48 (0.35)	0.57 (0.38)	0.24
monosemy	1.97 (2.43)	0.78 (0.32)	0.76 (0.35)	0.80 (0.31)	0.09

Table 3: Averages on nouns only (standard deviation in parentheses).

We create bins for each characterization, distinguishing four types of polysemy and three histogram bins for each of the other characteristics. Except for cosine to centroid, we use the distribution of all nouns recorded in the LDOCE dictionary (Proctor, 1978) to divide candidates across bins. For each characterization, we randomly draw examples from each bin until we reach a certain predefined number of examples for probably positive, probably negative or undecided candidates. The resulting distributions are summarized in Table 4.

We aim to include different types of ambiguity, since ambiguity is of particular interest for our further research. We are not aware of a lexical resource providing fine-grained information about types of ambiguity. To approximate it, we exploit metaphor annotations in the MIPVU corpus (Steen, 2010) and the distinction between homonymy, polysemy and monosemy information in the LDOCE dictionary. The third group we distinguish consists of other forms of polymsemy (metonymy, specialization and generalization). While it is not feasible to verify this approximation manually, we tested a number of tendencies which should hold if our approximation strategies are appropriate: The similarity between senses of ambiguous words should correlate with the semantic phenomena involved in it: Senses of homonymous words should be least similar while mentonymous senses should be most similar. This can be measured in terms of WordNet similarity or embedding vector similarity with the monosemous synset members of the senses. The sense similarity/distance can also be analyzed in terms of very broad semantic areas that a sense can fall into. Homonymous senses accidentally share the same form and metaphorical words often express mappings between abstract and concrete domains. Therefore, we expect that the latter two tend to have senses in both the abstract and concrete part of the WordNet hierarchy, while this should not be the case for metonymous senses (which typically remain restricted to one part of the hierarchy).

property	pos	neg	pos/neg	total
warm	20	28	118	166
hot	19	20	108	147
red	46	59	69	174
square	6	23	90	119
green	57	58	60	175
cold	18	22	81	121
sweet	28	1	145	174
blue	22	60	61	143
yellow	45	65	64	174
round	37	2	101	140
black	60	58	34	152
juicy	20	6	148	174
swim	57	61	62	180
roll	4	1	115	120
lay_eggs	61	61	32	154
fly	58	61	61	180
dangerous	63	61	17	141
used_in_cooking	59	60	60	179
female	57	11	48	116
wheels	54	16	45	115
wings	58	60	29	147
made_of_wood	59	12	81	152

Table 4: Overview of dataset size after sampling.

As the results summarized in Table 3 indicate, the ambiguity bins seem to provide a decent representation homonyms, words with metaphorical and metonymous senses and (for the same of comparison) monosemous words. We therefore use them for sampling.

4.5 Framework for Collecting Judgments

The resulting candidate concepts should be annotated in terms of their relations to the target property. To do this in an efficient way, we present

relation	examples	T/F
unusual	In an unusual situation, *chocolate* could be **pink**.	True
	In an unusual situation, *chocolate* could be brown.	False
affording_activity	Having **ink** is necessary for things a *pen* usually does or for things we usually do with a pen.	True
	Being **grey** is necessary for things a *car* usually does or for things we usually do with a *car*.	False
typical_of_concept	Being **spicy** is typical of a *chili pepper*.	True
	Being **sweet** is a typical property of a *carrot*.	False
variability_open	A *t-shirt* can be **white** or of another property of the same category as **white** there is a very wide set of possible options.	True
	A *pepper* can be **white** or of another property of the same category as *white* there is a very wide set of possible options.	False

Table 5: Examples of concept-property relations for crowd annotation with most appropriate True/False-judgment.

crowd workers with statements about the relation between a concept and a property and ask them to indicate whether it is generally true or false. We opt for this set up rather than presenting workers with all options, as it is faster and will most likely seem more attractive.[3] Rather than presenting generic, abstract descriptions of a property-concept pair, we present sentences such as the examples presented in Table 5, which are supposed to be natural-sounding and easy to judge.

5 Discussion and Conclusion

In this paper, we have outlined a method to create a dataset of semantic properties of concepts which can be used to evaluate whether and to what extent distributional models reflects semantic properties. This work can be positioned in our larger research goals, which involve creating transparent, interpretable lexical semantic representation in terms of semantic properties which lend themselves well for reasoning over ambiguity and variation. The dataset will be made available upon completion.[4]

The main goal of this paper is to propose a design for a dataset that can be used to test the ability of word embeddings to represent semantic properties. A more precise understanding of what information word embeddings can provide is highly relevant for improving NLP systems relying on embeddings as lexical semantic representations. Moreover, it can help in deciding whether embeddings are an appropriate representation in computational models of cognitive processes (as for instance discussed by Utsumi (2011)). Eventually, we plan to move towards data-derived interpretable word representations in terms of semantic properties.

The dataset proposed here enables us to use methods suggested in the area of studying representations and learning processes in neural networks, specifically diagnostic classification to test whether embeddings represent properties. In particular, we can go beyond the approach presented by Derby et al. (2018), who use all concepts for which a property has not been elicited as negative examples of a property.

In addition to proposing a dataset design, we offer specific hypotheses based on a variety of observations from different fields about information that is likely or unlikely to be expressed in English natural language corpora. Rather than making claims based on entire categories of semantic properties, we base our predictions on underlying factors involved in the relations between concepts and properties. By testing these hypotheses, we hope to go beyond insights from experimental approaches comparing the information captured in embeddings to semantic feature norm sets (e.g. Fagarasan et al. (2015), Herbelot and Vecchi (2015), Tsvetkov et al. (2015), Derby et al. (2018), Sommerauer and Fokkens (2018)).

Finally, we hope that comparing the relations captured by our dataset to traditional, taxonomic categories represented in WordNet may yield insights about the relation between properties of concepts and categorization. This could be extended to other languages to enable cross-linguistic comparisons.

[3] At this point, the exact set-up of the task is still under development. The resulting dataset will be made available once data have been collected.

[4] https://github.com/cltl/semantic_property_dataset

Acknowledgments

This research is funded by the PhD in the Humanities Grant provided by the Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) PGW.17.041 awarded to Pia Sommerauer and NWO VENI grant 275-89-029 awarded to Antske Fokkens. We would like to thank Emily Bender and anonymous reviewers for feedback that helped improve this paper. All remaining errors are our own.

References

Anja Arts, Alfons Maes, Leo Noordman, and Carel Jansen. 2011. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 1–10.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. 2018. Representation of word meaning in the intermediate projection layer of a neural language model. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 362–364.

Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 52–57.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

James J Gibson. 1954. The visual perception of objective motion and subjective movement. *Psychological Review*, 61(5):304.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Arthur M Glenberg and David A Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401.

Arthur M Glenberg. 1997. What memory is for. *Behavioral and brain sciences*, 20(1):1–19.

HP Grice. 1975. Logic and conversation. *Foundations of Cognitive Psychology*, page 719.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

P Proctor. 1978. *Longman Dictionary of Contemporary English*. Longman Group, Essex, UK.

Paula Rubio-Fernández. 2016. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, 7:153.

Julie C Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, 32(1):3–23.

Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054, Lisbon, Portugal.

Akira Utsumi. 2011. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive science*, 35(2):251–296.

Tony Veale and Yanfen Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

Tony Veale. 2013. The agile cliché: using flexible stereotypes as building blocks in the construction of an affective lexicon. In *New Trends of Research in Ontologies and Lexical Resources*, pages 257–275. Springer.

A Framework of semantic relations between concepts and properties

factor	relation description	example	represented	instances
impliedness	(A/an) [concept] is part of a larger category of which all members are [attribute].	**animate** - *cat*	no	all/most
typicality (of the concept)	(A/n) [concept] is a typical examples of things which are [attribute].	**green** - *broccoli*	no	all/most
typicality (of the property)	[attribute] is a typical property of (a/an) [concept].	**blue** - *sky*	yes	all/most
afforded (attribute)	Being [attribute] is necessary for activities/processes (a/an) [concept] is usually involved in.	**has a point** - *dagger*	yes	most/all
variability (distinction)	[attribute] is an important factor to distinguish different subcategories of members of the category [concept].	**grey** - *bear*	yes	some
variability (limited)	(A/an) [concept] can be [attribute] or another attribute of same category as [attribute] - there is a limited set of possible options.	**red** - *pepper*	yes	some
	(A/an) [concept] can be [attribute] or a bit more [attribute] or the opposite of [attribute].	**warm** - *water*	yes	some
variability (open)	(A/n) [concept] can be [attribute] or another attribute of same category as [attribute] - there is a very wide set of options.	**pink** - *t-shirt*	no	some
Variability (unlikely)	(A/an) [concept] is [attribute] could only be true in a rather unusual situation.	**blue** - *horse*	no	few/none
Variability (creative)	(A/an) [concept] is [attribute] can only be true in a creative, figurative way of speaking.	**round** - *idea*	no	few/none
Impossible	It is impossible that (a/an) [concept] is [attribute].	**solid** - *steam*	no	none

Table 6: Overview of relations between attributes and concepts.

factor	relation description	example	represented	instances
impliedness	(A/an) [concept] is/are part of a larger category of which all members can do/are involved in [activity].	**breathe** - *cat*	no	most/all
typicality (of the concept)	'[activity]' is a typical activity or process of (a/an) [concept].	**fly** - *bird*	no	all/most
typicality (of the property)	(A/an) [concept] is/are typical example(s) of things which do/are involved in the activity or process '[activity]'.	**hunt** - *tiger*	yes	all/most
afforded (activity)	(A/an) [concept] usually does/is involved in the activity or process '[activity]'.	**run** - *horse*	yes	most/all
	(A/an) [concept] can do/be involved in the activity or process '[activity]' but this is not what it/they usually does/do.	**roll** - *pen*	no	most/all
variability (distinction)	Doing/being involved in the activity or process '[activity]' is an important factor for distinguishing different subcategories of members of the category [concept].	**cooking** - *knife*	yes	some
variability (open)	(A/an) [concept] can do/be involved in the activity or process '[activity]' or not, but this is not an important factor for distinguishing different subcategories of members of the category [concept].	**play** - *dog*	no	some
Variability (unlikely)	(A/an) [concept] does/is involved in the activity or process '[activity]' could only be true in a highly unusual situation.	**fly** - *car*	no	few/none
Variability (creative)	(A/an) [concept] does/is involved in the activity or process '[activity]' can be only true in a creative, figurative way of speaking.	**fly** - *idea*	no	few/none
Impossible	It is impossible that (A/an) [concept] does/is involved in the activity or process [activity].	**fly** - *horse*	no	none

Table 7: Overview of relations between activities and concepts.

B Overview of selected properties

property type	category	properties
attributes	perceptual	warm, hot, red, square, green, cold, sweet, blue, yellow, round, black, juicy
	parts	wheels, wings, made_of_wood
	complex	dangerous, found_in_seas, used_in_cooking, female
activities	swim, roll, lay_eggs, fly	

Table 8: Overview of properties currently included (open for expansion).